New peptides under the s(ORF)ace of the genome

Article (Accepted Version)

New peptides under the s(ORF)ace of the genome

J.I. Pueyo[1†], E.G. Magny[1†], and J.P. Couso[1*]

1: Brighton and Sussex Medical School, University of Sussex, United Kingdom
†: These authors contributed equally to this work
*: correspondence to j.p.couso@sussex.ac.uk (J.P. Couso)

Keywords: smORFs, peptides, Ribosome Profiling, peptidomics, LncRNAs, uORFs.

Abstract

Hundreds of previously unidentified functional small peptides could exist in most genomes, but these sequences have been generally overlooked. The discovery of genes encoding small peptides with important functions in different organisms, has ignited the interest in these sequences, and led to an increasing amount of effort towards their identification.

Here, we review the advances, both, computational, and biochemical, that are leading the way in the discovery of putatively functional smORFs, as well as the functional studies that have been carried out as a consequence of these searches. The evidence suggests that smORFs form a substantial part of our genomes, and that their encoded peptides could have important functions in a variety of cellular functions.

Introduction

Deciphering the genetic information encoded in a genome is one of the main challenges in Biology. A constant improvement of sequencing and bioinformatics techniques has greatly advanced our understanding of this information but has also revealed the extent of its complexity. The difficulties associated with accurately predicting and annotating Small Open Reading Frame genes (smORFs) perfectly illustrate this complexity and the challenges it poses.

In the genome of most organisms there are hundreds of thousands of putative smORFs, consisting of a start-codon followed by in-frame codons and ending with a stop-codon [1-2]. Distinguishing translated and functional smORFs among this overwhelming and mostly spurious pool of sequences represents a major issue, which is particularly difficult to resolve since standard computational algorithms to identify coding sequences are generally not suited for small sequences [3-5]. Initially, short coding sequences (<100aa) were excluded from genome annotation pipelines [6], with the assumption that the majority of coding genes would code for larger proteins [7]. However, genes encoding small peptides have been identified in several organisms [8], like the *tarsal-less/polished rice/millepattes* gene, which codes for 11 aa-long peptides with important developmental functions in arthropods [9-12]. Such examples have led to the realisation that previously uncharacterised protein-coding smORFs with promising biological functions could exist in most genomes, and an increasing amount of effort has been directed towards their identification.

Here we will focus on the advances, both computational and biochemical that have been used to identify smORFs, and will present some of the different examples of smORFs which have been functionally characterised as a consequence of these studies.

Altogether, there is evidence suggesting that smORFs form a substantial part of our genomes and that their encoded peptides could be involved in a variety of cellular functions. Their characterisation could therefore lead to discoveries with important implications in cell biology and human health.

Systematic searches for putative coding smORFs using computational approaches.

Initial genome-wide searches for putative functional smORFs were conducted by bioinformatics methods designed to overcome the limitations of standard gene annotation algorithms. Generally, these methods were based on the analysis of sequence-composition frequencies (Figure 1A; see sORFfinder and CRITICA in BOX1), and/or on the evaluation of: **a)** the conservation of candidate smORF sequences in related species using pair-wise alignment-based tools (Figure 1B; BLAST [13]), and **b)** of their purifying selection (conservation of the aa relative to nt sequence) [14]. These initial studies identified several hundreds, and even thousands of putatively functional novel smORFs in the genomes of yeast, plants, flies, and mice [15-19], generally representing about 3-5% of the annotated genes in these organisms (Table 1).

In order to identify conserved coding sequences, more recent methods based on multiple sequence alignments incorporate phylogenetic distances and a model of nucleotide substitution rates, in the case of PhastCons [20], or a model of codon substitution frequencies, in the case of phyloCSF ([21], both built upon known coding and no-coding sequences (see Box 1). As shown below, these methods have sometimes been used

together with experimental methods in order to validate, or strengthen, the functionality of the smORFs identified as translated.

Ribosome Profiling: a biochemical approach for genome-wide translation assessment of smORFs

Next generation RNA sequencing (RNA-seq) has allowed to identify entire transcriptomes [22] and led to the unexpected realisation that a much higher than anticipated portion of the genome is transcribed (up to 85% in mammals [23] and 75% in flies [24]). A large proportion of these transcripts lack a "long" ORF of more than 100 aa, and have therefore been considered as long non-coding RNAs (LncRNAs), even though they otherwise resemble canonical mRNAs, having similar lengths, being transcribed by RNA-polymerase II, capped, poly-adenylated, and most even accumulating in the cytoplasm [25]. Although several LncRNAs have a well-established non-coding function [26], for the vast majority this remains unknown, making it plausible that some LncRNAs actually encode smORFs.

A method known as ribosome profiling (or Ribo-seq; Figure 1C) [27], consisting in sequencing nuclease-protected mRNA fragments (or footprints) bound by translating ribosomes (stabilized with an elongation inhibitor like cychlohexamide (CHX), allows to quantitatively and qualitatively measure the translation of these transcriptomes [28].

Different ribosome profiling studies, in a wide variety of species [29-40] have found that translation occurs in an almost pervasive fashion, detecting ribosome footprints in LncRNAs, in the untranslated regions (UTRs) of annotated transcripts, either upstream (uORFs) or

downstream (dORFs) of the CDS, and even overlapping the CDS of canonical mRNAs, with the vast majority of these corresponding to short ORFs (Table 1, and Table S1).

However there is some ambiguity with this method, since a ribosome bound fragment (RBF) read does not always strictly equate to an actively translated RNA fragment; a fragment of similar size could be obtained by a scanning ribosome, or other RNA-binding proteins [28]. Ribo-seq studies therefore employ different experimental or computational strategies to identify more accurately actively translated regions, involving the use of different metrics, such as RBF coverage, translation efficiency (TE: the ratio of RBFs / total mRNA reads), ribosomal release score (RRS), or codon phasing (see BOX1, [2]). Translation inhibitors, such as harringtonine (HR), which generates a pile-up of RBFs at the start codon, have also been used to identify translation initiation sites in actively translated ORFs [30]. Some of these studies have focused exclusively on the identification of translated smORFs in fruit flies [32], zebrafish [31], yeast [36], and mice [41].

In *Drosophila,* Aspden *et al.*[32] incorporated polysomal fractionation before ribo-seq to isolate cytoplasmic RNAs bound by 2-6 ribosomes and therefore actively translated, rather than those being scanned by single non-productive ribosomes or other RNA-binding proteins, this also enriched for RNAs encoding short ORFs (6 ribosomes being the maximum number that could fit in a 300 nt ORF). Using stringent RFB density and coverage thresholds, they corroborated the translation of 83% of the annotated smORFs transcribed in *Drosophila* S2 cells (228 out of 274), and found 2,708 and 313 novel translated smORFs in 5'UTRs and LncRNAs (Figure 2). Annotated smORFs, were found to be longer (~80 aa median) and with similar levels of "functionality" as canonical coding genes (conservation, aa usage and secondary structures) whereas the smORFs detected in 5'UTRs and lncRNAs,

were shorter (~20aa media length), and lacked the functional signatures observed in longer smORFs. However, some of these 5'UTR and LncRNA smORFs could be detected in epitope tagging experiments, displaying a similar sub-cellular localisation as canonical proteins, suggesting that some of them may encode functional peptides.

Bazzini *et al.* [31] performed Ribo-seq in zebrafish embryos, using ORFscore (See Box 1), a method which quantifies the 3-codon periodicity of the distribution of RBFs relative to the predicted ORF (phasing), a feature consistent with those ORFs being actively translated. Using this method, they validated the translation of 302 (52%) previously annotated smORFs, and identified 190 novel smORFs in previously uncharacterised transcripts and LncRNAs, as well as 311 uORFs and 93 ORFs in 3'UTRs (Figure 2; Table 1). In parallel, 63 novel smORFs were found using a conservation-based computational pipeline (micPDP) (see Box 1) in a catalogue of non-coding transcripts, 23 of them were also deemed translated by Ribo-seq, representing a pool of peptides highly likely to be translated and functional in zebrafish.

In yeast, 1,088 previously uncharacterised transcripts associated with poly-ribosomes (supporting their translation) were found [36]. Ribo-seq identified 185 of these as having sufficient footprint coverage, and TE scores to support smORF translation. Furthermore, 61 out of 80 transcripts from this pool, showed a codon triplet phasing bias to a single frame, suggesting their translation. Finally, 39 of these translated smORFs also showed varying extents of conservation among divergent yeast species, implying that they could be functional (Table 1).

Crappe *et al*.[41], combined a computational smORF search with Ribo-seq data to identify potentially coding smORFs in the mouse genome. A systematic search for potential coding

smORFs, conserved across mammalian species, was performed using sORFinder [16] and PhastCons (Box 1). Subsequently, a Support Vector Machine (SVM) learning algorithm, trained with sets of putatively non-coding and coding sequences, was used to classify the predicted smORFs, leading to the identification of some 40,000 smORFs with high coding probability in intergenic regions and LncRNAs. Independently , they re-analysed available ribosome profiling data from a mouse embryonic cell line [30], to identify translated smORFs (passing a coverage threshold, and showing a pile-up at their start codon when treated with HR). They identified 528 intergenic smORFs and 226 smORFs in LncRNAs, of these 401 and 89, respectively, were also found in the computational pipeline, representing a pool of smORFs likely to encode functional peptides (Table 1).

This study highlights the discrepancy in numbers that can exist between computational predictions and experimental detection. Part of this discrepancy could be explained by a possible high false positive rate in the bioinformatic pipeline, which could be due, for example, to the presence of conserved elements such as transposons, pseudogenes, and simple repeats [42]. It could also be explained by the fact that computational pipelines can search whole genomes for putative smORFs, whereas only the smORFs within transcripts expressed above a certain threshold in specific cells or tissues studied will be tested in Ribo-seq (or HPLC-MS) experiments.

Lee *et al.*[33]used a similar method, treating human and mouse cell lines with lactimidomycin (LTM), another initiation phase inhibitor. In this study they identified 227 annotated Human smORFs as translated (out of 694 annotated smORFs in ENSEMBL), as well as 288 ORFs in LncRNAs and 1,194 uORFs (most of them <100 aa long) (Figure 2; Table 1).

Altogether these studies show that thousands of smORFs are translated in eukaryotic genomes, with a substantial portion showing conservation and coding potential features, suggesting that a large repertoire of functional, yet uncharacterized peptides could exist in these organisms.

Detection of smORF peptides by mass spectrometry

The high-performance Liquid chromatography Mass-spectrometry (HPLC-MS) proteomics approach [43] has also been adapted to identify small peptides, mainly, by modifying the protocols for data analysis: instead of comparing candidate peptide spectrum matches (PSMs) to databases of annotated proteins, these are compared to databases generated *de novo*, based on all the possible translations of a given transcriptome (Figure 1D). Furthermore, standard proteomics require protein sequences to be supported by multiple PSMs. Because smORFs are too short to fit more than one PSM, this single PSM is usually required to pass the most stringent criteria in order to be unambiguously assigned to that smORF, potentially leading to a higher rate of false negatives.

Slavoff *et al.*[44] developed a peptidomics strategy, also applying specific experimental optimizations: inhibiting proteolysis, arguing that the proteolytic fragments of canonical proteins greatly increases the complexity of the peptidome and deteriorates the signal to noise ratio when it comes to identifying short peptides (themselves more susceptible to protease degradation), and using electrostatic-repulsion hydrophilic interaction chromatography (ERLIC) prior to HPLC-MS. They identified 86 novel peptides in human cells: 33 of them mapping to alternative CDS' in annotated transcripts (corresponding to uORFs,

dORFs, and smORFs overlapping annotated CDS'), 8 mapping to LncRNAs, and 49 of them mapping to previously un-annotated transcripts (Table 1).

This method was tested against other workflows [45], leading to two important observations: first, the use of ERLIC fractionation greatly increases the number of peptides detected (~10 fold) and second, there is an important lack of overlap between the peptides identified by different workflows, and even by different technical repeats, highlighting the stochastic nature of this technique, and the requirement of several repeats to achieve an optimal sampling saturation of the peptidome. In total, they analysed 3 different cell lines and a tumor sample, and identified a total of 311 short peptides, of which 237 are novel, with ~80% of them mapping to previously unannotated transcripts (Table 1), and the rest to alternative CDS within annotated transcripts with a similar distribution, in UTRs and overlapping CDS', as found by Slavoff *et al*.[44].

Another study, identified 1,259 novel peptides [46], by matching the spectra of different HPLC-MS data-sets (16 in total, covering a range of different human samples) to a custom database of predicted ORFs within annotated transcripts (mapping to UTRs and overlapping CDS'), suggesting that the translation of these "alternative" smORFs could be a wide-spread phenomenon. Interestingly, the majority of these peptides were identified in plasma and serum samples (1,118 / 1,259), implying that they could be secreted, although the reason or mechanism leading to this remains unknown (Table S1). Again, given the stochastic nature of this technique, this seemingly high number of identified novel peptides could be explained, in part, to the large number of samples analysed in this study.

Some of the Ribo-seq-based studies covered above have used HPLC-MS in order to validate their results (Table 1). In general, previously annotated smORFs tend to be more abundantly

detected by HPLC-MS than uORFs or LncRNA smORFs; in Aspden *et al.*[32] and Bazzini *et al.*[31] detected almost a third of the 228 and 302 annotated translated smORFs, but Aspden *et al.*[32] failed to identify any peptide from LncRNAs or uORFs, and Bazzini *et al.* [31] only identified 3 and 17, respectively. Similarly, only a handful of peptides corresponding to uORFs and LncRNas have been detected by HPLC-MS in studies that detected hundreds by Ribo-seq in humans (Figure 2; Table 1). These results clearly highlight a difference of sensitivity between these methods, and could also be in agreement with the segregation of these smORFs into two different functional classes, as observed by Aspden *et al.*[32], with smaller uORF and LncRNA ORFs showing lower conservation features and being less likely to be translated into functional peptides than longer annotated smORFs. These results could also be explained by the stochastic nature of the peptidomics method observed by Slavoff *et al*.[44], with the peptides from LncRNAs or uORFs being generally smaller, and therefore probably less stable, less abundant, and having lower chances of being detected. In that sense, detection by peptidomics could be considered as a convincing proof of translation, and as an indication of probable function, but the opposite is not necessarily true (Figure 2). Also, it is important to point out that these studies did not use the extensively optimized protocols (with proteolysis free conditions, ERLIC fractionation, and multiple technical repeats), which may have improved the detection of these smaller peptides.

Computational and biochemical strategies unravel novel smORF peptide functions

Although these computational and biochemical approaches have identified hundreds of translated and conserved smORFs, previous systematic functional studies (based on random

mutagenesis) in different organisms, have failed to find them. This disparity could be explained by the lower probability of mutagens to target a small ORF in comparison to larger canonical ones. In addition, and as exemplified by several of the examples covered below, these small peptides may act as regulators of cellular processes requiring a very specific and in-depth phenotypic analysis, in order to detect mutants. As a result, only a handful of smORFs, found serendipitously, had been functionally characterised prior to these extensive smORF searches [8].

However, these genome-wide smORF searches have aided the functional characterisation of smORFs by identification of putative candidates. Following their bioinformatic prediction, some studies have carried out high-throughput smORF functional screens in yeast, [15] and in plants [47], and found dozens of functional smORFs, with several being essential (Table 1).

Other studies have focused on a more in-depth characterisation of specific smORFs. One example is the *Sarcolamban (Scl)* gene [48], previously annotated as a non-coding gene [49], but identified by a bioinformatics approach as a potential functional smORF in *Drosophila* (Table 1;[17]). *Scl* encodes for two 28 and 29aa transmembrane related-micropeptides which act as inhibitors of SERCA calcium pump and regulate heart muscle contraction (Figure 2A;[48]). Importantly, these peptides appear to be functional homologues to the vertebrate Sarcolipin and Phospholamban peptides, thereby uncovering an ancestral family of smORFs conserved from insects to humans [48]. More recently, another member of this family, Myoregulin (46 aa) [50], and a novel small peptide, DWORF (34 aa) [51] with an antagonistic function (since it enhances the activity of SERCA), have both been identified in mice from previously non-coding annotated transcripts. Toghether with Sln and Pln, (and Scl

in flies) these peptides contribute to the smORF-based regulatory repertoire that regulates calcium dynamics and seemingly participate in conferring different muscles with specific contractility properties [50].

Another example is the *toddler/apela* gene, which was identified in a ribo-seq-based search for novel signalling peptides in zebrafish (Table 1;[52]). The *apela* gene encodes a secreted 58 aa peptide, that binds to the Apelin receptor and promotes cell mobility during gastrulation [52]. This novel peptide also shows a great extent of conservation across vertebrates.

Similarly, the *Drosophila hemotin (hemo)* gene was identified as a putative functional smORF by a computational study [17], and subsequently, its translation supported by ribosome profiling and proteomics studies [32, 53](Table 1). *hemo* is expressed in hemocytes (*Drosophila* macrophages) where it regulates endosomal maturation, and phagocytosis, by inhibiting the activity of phosphotidyl-inositol kinases through an interaction with 14-3-3z (Figure 2B;[54]). Interestingly, this regulatory mechanism also appears to have been conserved across evolution as the vertebrate Stannin (Snn), a factor involved in organometallic cytotoxicity [55], is the functional Hemo homologue in flies and mouse macrophages [54].

In humans, the MRI-2 peptide, which was shown to stimulate double-strand break repair through a direct interaction with the DNA end binding protein *Ku* (Figure 2C;[56]), was functionally characterised because it appeared as translated in a HPLC-MS screen for novel short peptides (Table 1;[44]).

These examples highlight the contribution of these bioinformatic and experimental approaches in the identification of functional smORFs. They also strengthen our view about the complexity and biological relevance of these peptides, which can regulate a diversity of cellular processes, with their function being conserved, in some cases, across vast evolutionary distances. Overall, their study can certainly have important implications in cell biology, and even in human medical research[57].

Concluding remarks and future perspectives

In this review, we have shown that there is extensive evidence supporting the translation of substantial numbers of smORFs in a variety of organisms. This evidence is likely to increase as new methods and metrics are developed to analyse ribo-seq data more robustly in order to identify *bona fide* translated regions (BOX 2). For example, Ingolia *et al.*[58] recently developed a metric, based on assessing the distribution of RPF lengths (Fragment Length Organisation Similarity Score or FLOSS) which can accurately distinguish between reads protected by the translation-engaged 80s ribosomal conformation, from reads obtained from the protection of other non-translating ribosomal conformations (40s and 60s), or other RNA-binding proteins. Other groups have used classification algorithms, such as the random forest-based Translated ORF classifier (TOC) [29, 52], the logistic regression-based ORF-rater [59], or the SVM-based RibORF [60], which integrate different Ribo-seq metrics, and their profiles on known coding and non-coding regions, to identify translated ORFs. All these studies support the translation of hundreds of novel small peptides encoded in transcripts previously thought to be non-coding and as uORFs, in vertebrates.

It remains challenging, however, to identify which among this ever-growing set of Ribo-seq-supported translated smORFs encode functional peptides, from those representing "translational noise", or acting as translation-dependent regulatory sequences. Abundant evidence supports the role of uORFs as translational regulators, exerting this function through their engagement of ribosomes [61-63], inferring this is the main, canonical function of uORFs. Similarly, it has been suggested that smORFs within LncRNAs, or overlapping annotated coding mRNAs, could function mainly as regulators of transcript stability, by engaging the non-sense mediated decay (NMD) pathway [64-65] . Nonetheless, as shown above, several smORF-encoded peptides with important functions have been identified in previously non-coding RNAs proving that these sequences can certainly encode functional peptides [10, 48, 50-52, 54]. There are even examples of canonical non-coding RNAs, such as pri-miRNAs [66] and ribosomal RNAs [67-69], encoding biologically active peptides with well characterised functions. Similarly some uORFs have been shown to exert their regulatory function through their encoded peptides, with this regulation depending on their aa sequence [70], and being able to occur in *trans* [71-74].

To assess the functional potential of smORFs, some studies have used an integrative approach to take advantage of the extensive RNA-seq, Ribo-seq, and HPLC-MS datasets available, to assess the translation, conservation, and coding potential of smORFs in several organisms. Mackoviak *et al.*[75] identified, computationally, a total of 2,002 novel putatively functional smORFs in 5 different organisms, based on their conservation patterns (obtained, briefly, with an SVM-based classifier, taking into account ORF conservation in multiple alignments, and PhyloCSF and PhastCons scores). These peptides map mostly to UTRs and LncRNAs, show little homology to known proteins, and are shorter than annotated smORFs,

also having different aa sequence properties [75]. Interestingly these smORFs have Ribo-seq ORFscore values that are higher than non-coding controls, but lower than annotated smORFs. Similarly, Ruiz-Orera *et al.*[76] found that, in several species, smORFs in lncRNAs have intermediate Ribo-seq and conservation features, which resemble those of newly evolved peptides. These results are, overall, reminiscent to those of Aspden *et al.* [32] in *Drosophila*, reinforcing the idea of functionally distinct classes of smORFs.

Although these studies provide valuable information regarding the functional potential of smORFs, they remain elusive about their specific functions. These systematic smORF searches have the ultimate aim of advancing genome annotations, which ultimately entails the attribution of specific functions to these newly detected smORFs, and this functional characterisation certainly poses the next challenge towards which an increased amount of efforts should be directed. Advances in gene editing technologies such as CRIPSR, which allow to relatively quickly generate specific mutants in most organisms [77], and the development of more sensitive phenotypical screens and biochemical assays to accurately assess peptide functions[57], will help to start filling this void of functional information in the genome.

References:

1. Basrai, M.A., P. Hieter, and J.D. Boeke, *Small Open Reading Frames: Beautiful Needles in the Haystack.* Genome Research, 1997. **7**(8): p. 768-771.
2. Mumtaz, M.A. and J.P. Couso, *Ribosomal profiling adds new coding sequences to the proteome.* Biochem Soc Trans, 2015. **43**(6): p. 1271-6.
3. Wang, J., et al., *Vertebrate gene predictions and the problem of large genes.* Nat Rev Genet, 2003. **4**(9): p. 741-9.
4. Cheng, H., et al., *Small Open Reading Frames: Current Prediction Techniques and Future Prospect.* Curr Protein Pept Sci, 2011.
5. Sleator, R.D., *An overview of the current status of eukaryote gene prediction strategies.* Gene, 2010. **461**(1-2): p. 1-4.
6. Goffeau, A., et al., *Life with 6000 genes.* Science, 1996. **274**(5287): p. 546, 563-7.
7. Dujon, B., et al., *Complete DNA sequence of yeast chromosome XI.* Nature, 1994. **369**(6479): p. 371-8.
8. Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**: p. 193–204.
9. Savard, J., et al., *A segmentation gene in Tribollium produces a polycistronic mRNA that codes for multiple conserved peptides.* Cell, 2006. **126**(3): p. 559-569.
10. Galindo, M.I., et al., *Peptides encoded by short ORFs control development and define a new eukaryotic gene family.* PLoS Biology, 2007. **5**(5): p. 1052-1062.
11. Kondo, T., et al., *Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA.* Nature Cell Biology, 2007. **9**(6): p. 660-U87.
12. Zanet, J., et al., *Pri sORF peptides induce selective proteasome-mediated protein processing.* Science, 2015. **349**(6254): p. 1356-8.
13. Altschul, S.F., et al., *Basic local alignment search tool.* Journal Of Molecular Biology, 1990. **215**(3): p. 403-410.
14. Nekrutenko, A., K.D. Makova, and W.H. Li, *The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study.* Genome Res, 2002. **12**(1): p. 198-202.
15. Kastenmayer, J.P., et al., *Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae.* Genome Res., 2006. **16**(3): p. 365-373.
16. Hanada, K., et al., *A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection.* Genome Research, 2007. **17**(5): p. 632-640.
17. Ladoukakis, E., et al., *Hundreds of putatively functional small open reading frames in Drosophila.* Genome Biol, 2011. **12**(11): p. R118.
18. Frith, M.C., et al., *The abundance of short proteins in the mammalian proteome.* Plos Genetics, 2006. **2**(4): p. 515-528.
19. Kessler, M.M., et al., *Systematic Discovery of New Genes in the Saccharomyces cerevisiae Genome.* Genome Res., 2003. **13**(2): p. 264-271.
20. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
21. Lin, M.F., I. Jungreis, and M. Kellis, *PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.* Bioinformatics, 2011. **27**(13): p. i275-82.
22. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.
23. Hangauer, M.J., I.W. Vaughn, and M.T. McManus, *Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs.* PLoS Genet, 2013. **9**(6): p. e1003569.

24. Graveley, B.R., et al., *The developmental transcriptome of Drosophila melanogaster.* Nature, 2011. **471**(7339): p. 473-9.
25. van Heesch, S., et al., *Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes.* Genome Biol, 2014. **15**(1): p. R6.
26. Kung, J.T., D. Colognori, and J.T. Lee, *Long noncoding RNAs: past, present, and future.* Genetics, 2013. **193**(3): p. 651-69.
27. Ingolia, N.T., et al., *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* Science, 2009. **324**(5924): p. 218-23.
28. Ingolia, N.T., *Ribosome profiling: new views of translation, from single codons to genome scale.* Nat Rev Genet, 2014. **15**(3): p. 205-213.
29. Chew, G.L., et al., *Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs.* Development, 2013. **140**(13): p. 2828-34.
30. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of Mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.
31. Bazzini, A.A., et al. (2014) *Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation*. EMBO J, embj.201488411 DOI: embj.201488411 [pii]
10.1002/embj.201488411.
32. Aspden, J.L., et al., *Extensive translation of small ORFs revealed by polysomal ribo-Seq* eLife, 2014. **3**: p. e03528.
33. Lee, S., et al., *Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution.* Proc Natl Acad Sci U S A, 2012. **109**(37): p. E2424-32.
34. Duncan, C.D. and J. Mata, *The translational landscape of fission-yeast meiosis and sporulation.* Nat Struct Mol Biol, 2014. **21**(7): p. 641-7.
35. Vasquez, J.J., et al., *Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages.* Nucleic Acids Res, 2014. **42**(6): p. 3623-37.
36. Smith, J.E., et al., *Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae.* Cell Rep, 2014. **7**(6): p. 1858-66.
37. Brar, G.A., et al., *High-resolution view of the yeast meiotic program revealed by ribosome profiling.* Science, 2012. **335**(6068): p. 552-7.
38. Stern-Ginossar, N., et al., *Decoding human cytomegalovirus.* Science, 2012. **338**(6110): p. 1088-93.
39. Dunn, J.G., et al., *Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.* eLife, 2013. **2**: p. e01179.
40. Juntawong, P., et al., *Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis.* Proc Natl Acad Sci U S A, 2014. **111**(1): p. E203-12.
41. Crappe, J., et al., *Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs.* BMC Genomics, 2013. **14**: p. 648.
42. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.
43. Shen, T.-L., *High-Performance Liquid Chromatography/Mass Spectrometry in Peptide and Protein Analysis*, in *Encyclopedia of Analytical Chemistry*. 2006, John Wiley & Sons, Ltd.
44. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.
45. Ma, J., et al., *Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue.* J Proteome Res, 2014. **13**(3): p. 1757-65.
46. Vanderperre, B., et al., *Direct detection of alternative open reading frames translation products in human significantly expands the proteome.* PLoS One, 2013. **8**(8): p. e70698.

47. Hanada, K., et al., *Small open reading frames associated with morphogenesis are hidden in plant genomes.* Proc Natl Acad Sci U S A, 2013. **110**(6): p. 2395-400.
48. Magny, E.G., et al., *Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames.* Science, 2013. **341**(6150): p. 1116-20.
49. Tupy, J.L., et al., *Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(15): p. 5495-5500.
50. Anderson, Douglas M., et al., *A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance.* Cell, 2015. **160**(4): p. 595-606.
51. Nelson, B.R., et al., *Muscle physiology. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle.* Science, 2016. **351**(6270): p. 271-5.
52. Pauli, A., et al., *Toddler: an embryonic signal that promotes cell movement via Apelin receptors.* Science, 2014. **343**(6172): p. 1248636.
53. Brunner, E., et al., *A high-quality catalog of the Drosophila melanogaster proteome.* Nat Biotechnol, 2007. **25**(5): p. 576-83.
54. Pueyo, J.I., et al., *Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans.* PloS Biology, 2016.
55. Billingsley, M.L., et al., *Functional and structural properties of stannin: roles in cellular growth, selective toxicity, and mitochondrial responses to injury.* J Cell Biochem, 2006. **98**(2): p. 243-50.
56. Slavoff, S.A., et al., *A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining.* J Biol Chem, 2014. **289**(16): p. 10950-7.
57. Saghatelian, A. and J.P. Couso, *Discovery and characterization of smORF-encoded bioactive polypeptides.* Nat Chem Biol, 2015. **11**(12): p. 909-16.
58. Ingolia, N.T., et al., *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.
59. Fields, A.P., et al., *A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation.* Mol Cell, 2015. **60**(5): p. 816-27.
60. Ji, Z., et al., *Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins.* Elife, 2015. **4**.
61. Calvo, S.E., D.J. Pagliarini, and V.K. Mootha, *Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans.* Proc Natl Acad Sci U S A, 2009. **106**(18): p. 7507-12.
62. Wang, X.Q. and J.A. Rothnagel, *5'-untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation.* Nucleic Acids Res, 2004. **32**(4): p. 1382-91.
63. Morris, D.R. and A.P. Geballe, *Upstream open reading frames as regulators of mRNA translation.* Mol Cell Biol, 2000. **20**(23): p. 8635-42.
64. Mendell, J.T., et al., *Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise.* Nat Genet, 2004. **36**(10): p. 1073-8.
65. Tani, H., M. Torimura, and N. Akimitsu, *The RNA Degradation Pathway Regulates the Function of GAS5 a Non-Coding RNA in Mammalian Cells.* PLoS ONE, 2013. **8**(1): p. e55684.
66. Lauressergues, D., et al., *Primary transcripts of microRNAs encode regulatory peptides.* Nature, 2015. **520**(7545): p. 90-3.
67. Guo, B., et al., *Humanin peptide suppresses apoptosis by interfering with Bax activation.* Nature, 2003. **423**(6938): p. 456-461.
68. Lee, C., et al., *The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance.* Cell Metab, 2015. **21**(3): p. 443-54.
69. Paharkova, V., et al., *Rat Humanin is encoded and translated in mitochondria and is localized to the mitochondrial compartment where it regulates ROS production.* Mol Cell Endocrinol, 2015. **413**: p. 96-100.

70. Ebina, I., et al., *Identification of novel Arabidopsis thaliana upstream open reading frames that control expression of the main coding sequences in a peptide sequence-dependent manner.* Nucleic Acids Research, 2015. **43**(3): p. 1562-1576.
71. Akimoto, C., et al., *Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites.* Biochim Biophys Acta, 2013. **1830**(3): p. 2728-38.
72. Nguyen, H.L., X. Yang, and C.J. Omiecinski, *Expression of a novel mRNA transcript for human microsomal epoxide hydrolase (EPHX1) is regulated by short open reading frames within its 5'-untranslated region.* Rna, 2013. **19**(6): p. 752-66.
73. Pendleton, L.C., et al., *Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading frame.* J Biol Chem, 2005. **280**(25): p. 24252-60.
74. Diba, F., C.S. Watson, and B. Gametchu, *5'UTR sequences of the glucocorticoid receptor 1A transcript encode a peptide associated with translational regulation of the glucocorticoid receptor.* J Cell Biochem, 2001. **81**(1): p. 149-61.
75. Mackowiak, S.D., et al., *Extensive identification and analysis of conserved small ORFs in animals.* Genome Biol, 2015. **16**: p. 179.
76. Ruiz-Orera, J., et al., *Long non-coding RNAs as a source of new peptides.* Elife, 2014. **3**.
77. Sander, J.D. and J.K. Joung, *CRISPR-Cas systems for editing, regulating and targeting genomes.* Nat Biotechnol, 2014. **32**(4): p. 347-55.
78. Guttman, M., et al., *Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.* Cell, 2013. **154**(1): p. 240-51.
79. Kong, L., et al., *CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.* Nucleic Acids Research, 2007. **35**(Web Server issue): p. W345-W349.
80. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.

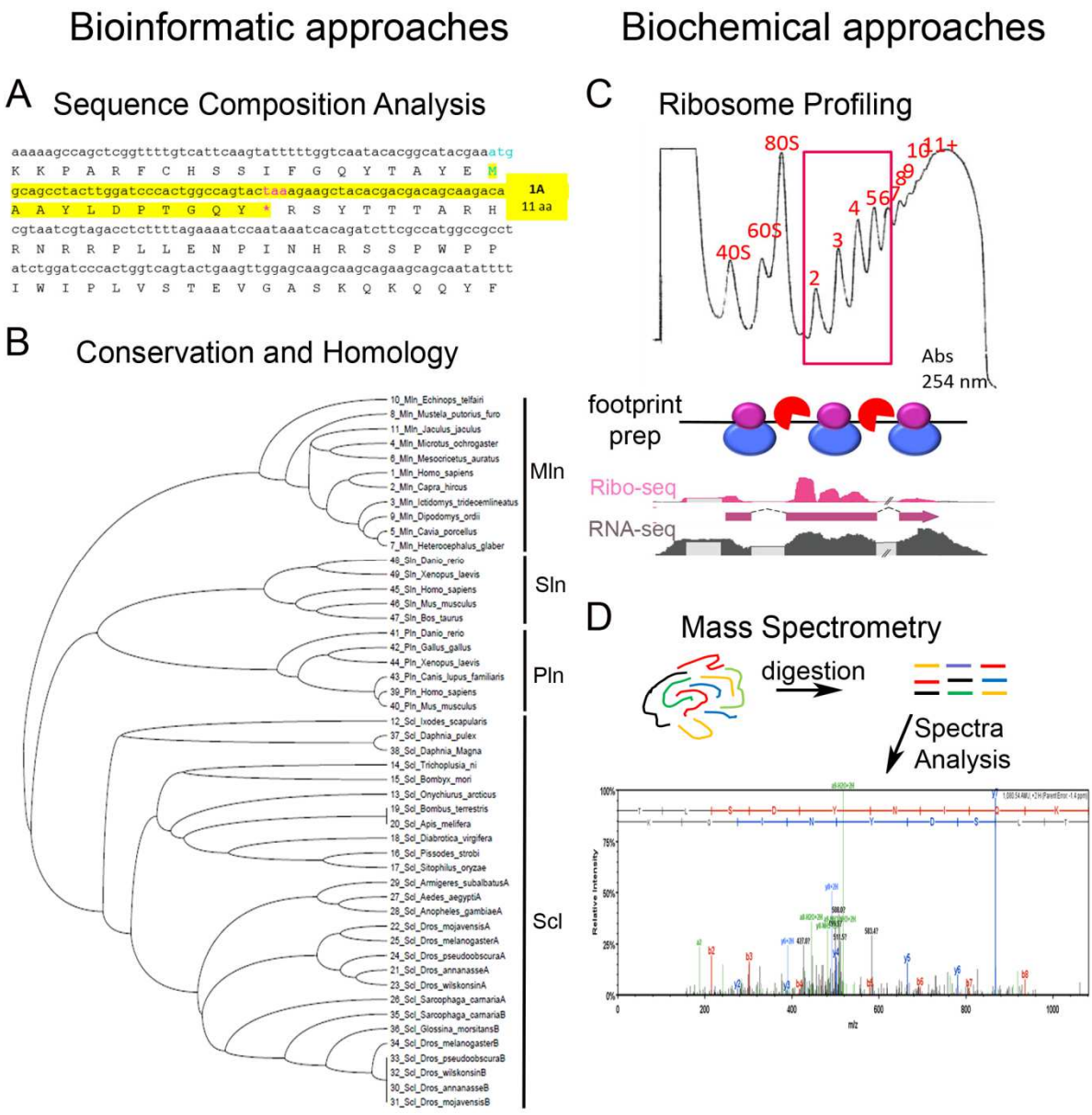
Bioinformatic approaches
Biochemical approaches
A
Sequence Composition Analysis
aaaaagccagctcggttttgtcattcaagtatttttggtcaatacacggcatacgaaatg
K K P A R F C H S S I F G Q Y T A Y E M
gcagcctacttggatcccactggccagtactaaagaagctacacgacgacagcaagaca
A A Y L D P T G Q Y * R S Y T T T A R H
1A
11 aa
cgtaatcgtagacctcttttagaaaatccaataaatcacagatcttcgccatggccgcct
R N R R P L L E N P I N H R S S P W P P
atctggatcccactggtcagtactgaagttggagcaagcaagcagaagcagcaatatttt
I W I P L V S T E V G A S K Q K Q Q Y F
B
Conservation and Homology
10_Mln_Echinops_telfairi
8_Mln_Mustela_putorius_furo
11_Mln_Jaculus_jaculus
4_Mln_Microtus_ochrogaster
6_Mln_Mesocricetus_auratus
1_Mln_Homo_sapiens
2_Mln_Capra_hircus
3_Mln_Ictidomys_tridecemlineatus
9_Mln_Dipodomys_ordii
5_Mln_Cavia_porcellus
7_Mln_Heterocephalus_glaber
48_Sln_Danio_rerio
49_Sln_Xenopus_laevis
45_Sln_Homo_sapiens
46_Sln_Mus_musculus
47_Sln_Bos_taurus
41_Pln_Danio_rerio
42_Pln_Gallus_gallus
44_Pln_Xenopus_laevis
43_Pln_Canis_lupus_familiaris
39_Pln_Homo_sapiens
40_Pln_Mus_musculus
12_Scl_Ixodes_scapularis
37_Scl_Daphnia_pulex
38_Scl_Daphnia_Magna
14_Scl_Trichoplusia_ni
15_Scl_Bombyx_mori
13_Scl_Onychiurus_arcticus
19_Scl_Bombus_terrestris
20_Scl_Apis_mellifera
18_Scl_Diabrotica_virgifera
16_Scl_Pissodes_strobi
17_Scl_Sitophilus_oryzae
29_Scl_Armigeres_subalbatusA
27_Scl_Aedes_aegyptiA
28_Scl_Anopheles_gambiaeA
22_Scl_Dros_mojavensisA
25_Scl_Dros_melanogasterA
24_Scl_Dros_pseudoobscuraA
21_Scl_Dros_annanasseA
23_Scl_Dros_wilskonsinA
26_Scl_Sarcophaga_carnariaA
35_Scl_Sarcophaga_carnariaB
36_Scl_Glossina_morsitansB
34_Scl_Dros_melanogasterB
33_Scl_Dros_pseudoobscuraB
32_Scl_Dros_wilskonsinB
30_Scl_Dros_annanasseB
31_Scl_Dros_mojavensisB
Mln
Sln
Pln
Scl
C
Ribosome Profiling
80S
60S
40S
2
3
4
5
6
7
8
9
10
11+
Abs
254 nm
footprint
prep
Ribo-seq
RNA-seq
D
Mass Spectrometry
digestion
Spectra
Analysis
Relative Intensity
m/z

FIGURE 1

Figure 1. Bioinformatic and biochemical approaches for the prediction of putative function smORFs.

Bioinformatic approaches: **A-** Nucleotide composition analyses of primary smORF sequences (tarsal-less 1A ORF;yellow), such as codon composition or hexamere nucleotide frequencies, are able to determine their coding potential, since the nucleotide frequencies of functional protein-coding ORFs are not random, due to a biased codon usage. **B-** Functional protein-coding sequences are under evolutionary constrains. Identification of smORF in closely related species allows to assess whether nucleotide changes are constrained to maintain the aa sequence (Ka/Ks). Furthermore, phylogenetic analyses of smORF homologues predict conserved motifs, or protein domains, which can be further used to identify distant homologues, as shown by the phylogenetic tree of Sarcolamban family.

Biochemical approaches: **C-** Ribosome profiling is based on sequencing of nuclease protected-ribosome bound RNA fragments (footprints), and allows a qualitative and quantitative genome-wide assessment of translation. Separation of polysomal fractions (red rectangle), enables the isolation of actively translated smORF transcripts (Poly-Ribo-Seq), and in combination with Ribo-seq, has detected translated smORFs. **D-** Mass spectrometry (MS) has detected smORF-encoded products from a digested protein sample by matching experimental spectra to predicted spectra from a reference/custom protein-database.

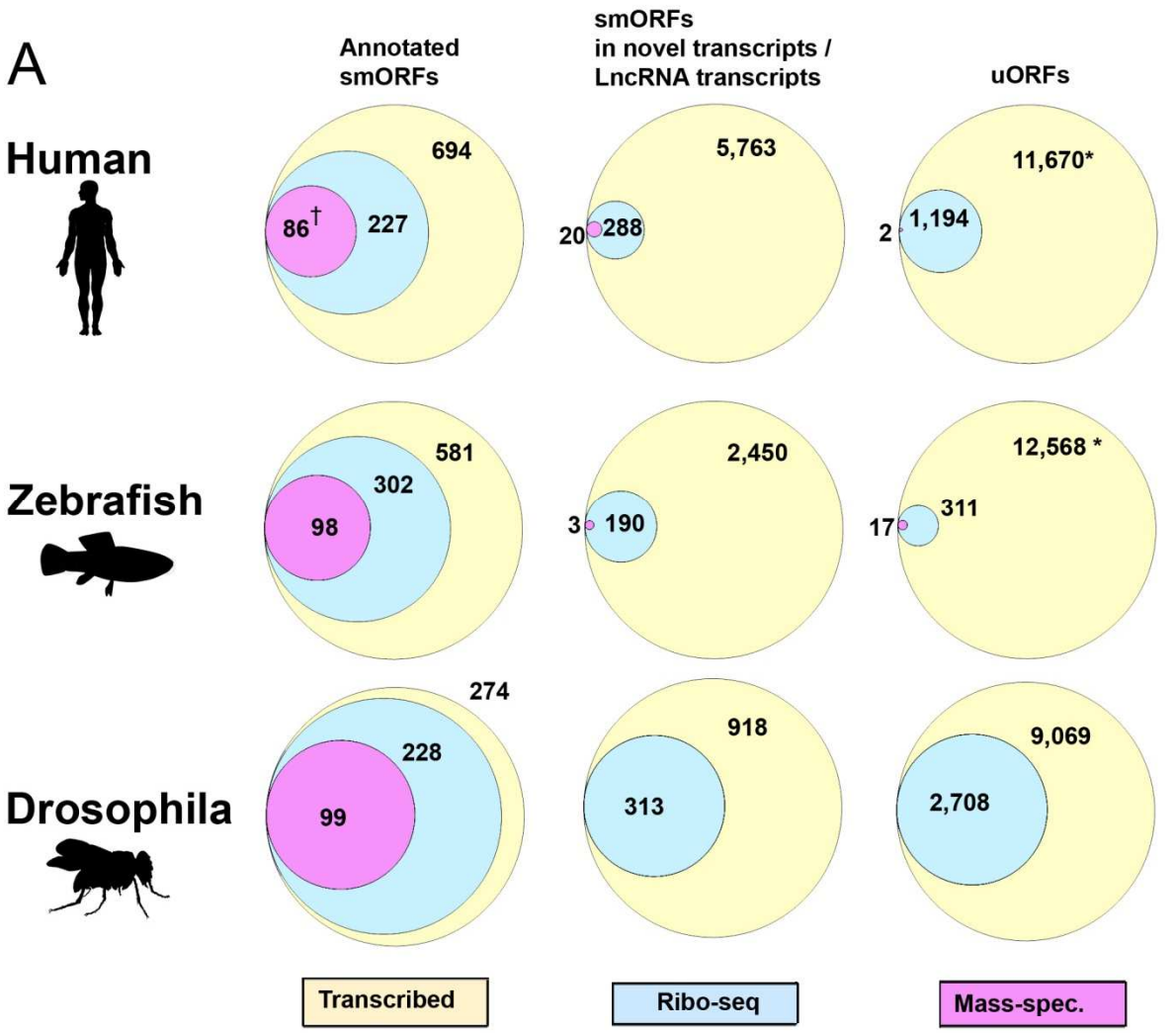
A
Annotated smORFs
smORFs in novel transcripts / LncRNA transcripts
uORFs
Human
694
86†
227
5,763
20
288
11,670*
2
1,194
Zebrafish
581
302
98
2,450
3
190
12,568 *
17
311
Drosophila
274
228
99
918
313
9,069
2,708
Transcribed
Ribo-seq
Mass-spec.
B
functional peptides ?
size, peptide stability, transcript expression, homology, protein domains
Mln
DWORF
MRI-2
Toddler
Scl
hemotin
tal
MKKS
EPHX1
NR3C1
ASS1

FIGURE 2

Figure 2. Different classes of smORFs detected by Ribo-seq and HPLC-MS in Humans, Zebrafish and fruit flies.

A- Venn diagrams representing the number of smORFs detected by Ribo-seq (blue) or HPLC-MS (pink), relative to the total number of transcripts encoding each class of smORF (yellow) in humans[33, 44], zebrafish [31]and fruit flies [32]. In these organisms, HPLC-MS detects very few peptides from LncRNAs and uORFs (0%-0.3%), compared to annotated smORFs (12-33%), whereas Ribo-seq still detects a substantial amount of LncRNA smORFs and uORFs (3-30%, compared to 30-80% annotated smORFs), highlighting the difference in sensitivity between these techniques. The number of transcribed uORFs (*) was inferred from the number of transcripts with uORFs identified in other studies, for humans [61] and for zebrafish [29]; the number of peptides identified in humans by HPLC-MS (†) were obtained from Mackowiak et al.[75]

B- The higher detection rates of annotated smORFs by HPLC-MS could be due to their higher levels of expression, and larger (and more stable) peptides, which also correlate with their closer resemblance to canonical proteins, in terms of functional signatures (protein domain content, conservation). Although these observations imply that annotated smORFs represent a functionally distinct class from LncRNA smORFs and uORFs, the identification of a growing number of biologically active peptides encoded in previously non-coding RNAs and uORFs (italics) proves that their functionality should not be systematically discarded.

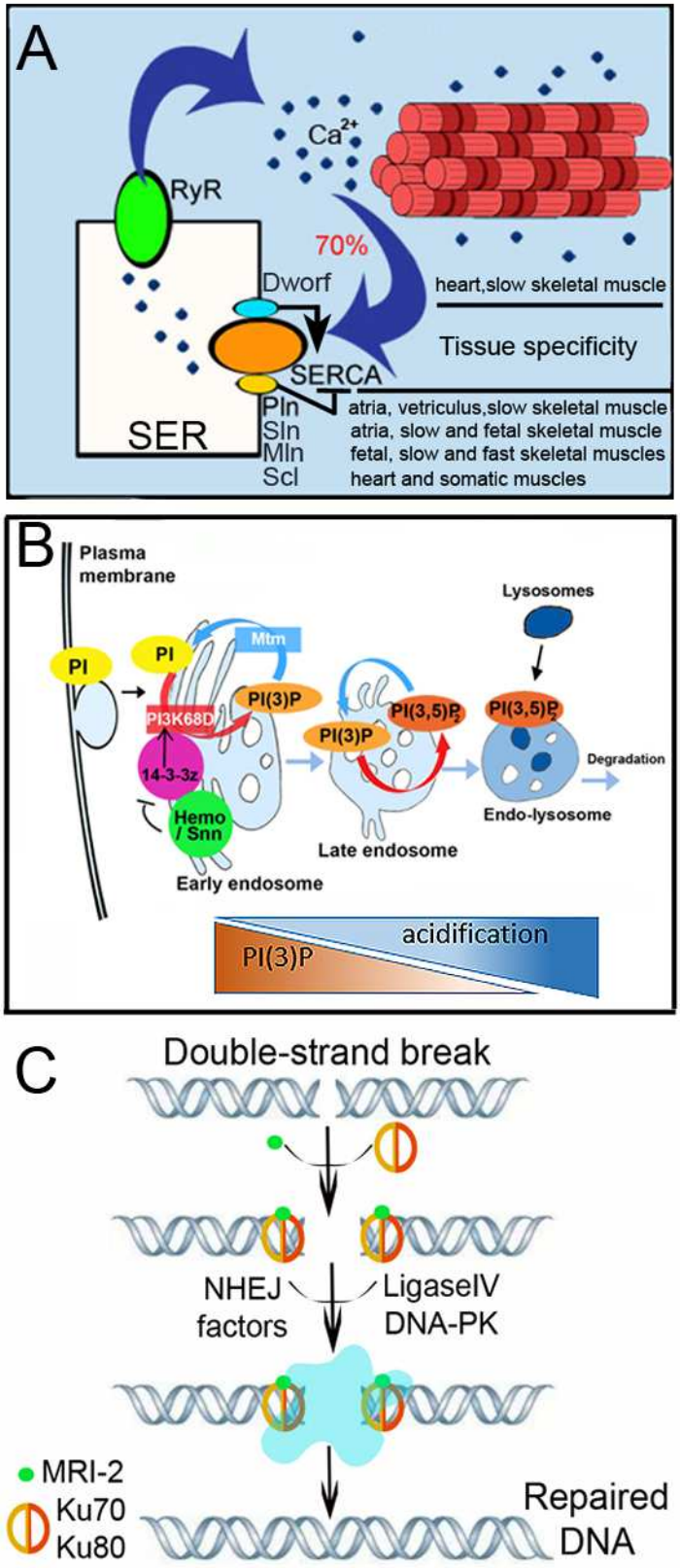
A
Ca2+
RyR
70%
Dworf
heart,slow skeletal muscle
Tissue specificity
SERCA
SER
Pln
Sln
Mln
Scl
atria, vetriculus,slow skeletal muscle
atria, slow and fetal skeletal muscle
fetal, slow and fast skeletal muscles
heart and somatic muscles
B
Plasma membrane
Lysosomes
PI
PI
Mtm
PI(3)P
PI3K68D
14-3-3z
Hemo / Snn
Early endosome
PI(3)P
PI(3,5)P2
Late endosome
PI(3,5)P2
Degradation
Endo-lysosome
acidification
PI(3)P
C
Double-strand break
NHEJ factors
LigaseIV
DNA-PK
MRI-2
Ku70
Ku80
Repaired DNA

FIGURE 3

Figure 3. Cellular functions of conserved smORF micropeptides.

A- Muscle performance depends on intracellular levels of Ca^{2+} regulated by the Ryanodine receptors (RyR) and Sarcoendoplasmic reticulum (SER) calcium ATPase (SERCA) pump. A conserved family of smORF peptides bind SERCA inhibiting its activity. Their members, Sarcolamban (Scl) in *Drosophila* and Sarcolipin (Sln) and Phospholamban (Pln) and Myoregulin (Mln) in vertebrates, display specific expression patterns. In addition, a new vertebrate smORF, DWORF, activates SERCA by competitively displacing SERCA inhibitors.

B-The *Hemotin (Hemo)/Stannin (Snn)* family is necessary for regulation of phagocytosis in *Drosophila* and mouse macrophages. Trafficking of phagocytised particles depends on the phosphorylation states of phosphatidyl-inositol (PI). At early endosomes, PI is phosphorylated into PI(3)P by the PI3Kinase (PI3K68D). At late endosomes PI(3)P is phosphorylated into PI(3,5)P2, which leads to lysosome fusion (acidification) and degradation of cargo. Vesicle trafficking can be reversed by PI(3,5)P2 dephosphorylation by Myotubulurin phosphatases (Mtm). Therefore, maturation of phagocytized particles correlates positively with acidification and negatively with PI(3)P. The 88aa-Hemo/Snn peptides inhibit a 14-3-3ζ-mediated Pi368Dkinase activation.

C- In humans, the MRI-2 peptide is involved in the non-homologous end joining (NHEJ) double-strand break (DSB) DNA repair. This 69aa-long peptide is recruited to the nucleus upon DSBs induction, where it binds Ku70/Ku80 heterodimers, and stimulates DNA ligation through NHEJ.

Table 1. The Identification of putative functional smORFs using computational and experimental approaches, has led to their functional characterisation.

smORF searches with functional outcomes

Organism:	Method/Metrics:	Indentified smORFs:	Comments:	Ref.	Functional outcomes:	Ref.
Yeast	Conservation, transcription, optimal codon adaptation index	588	Putative functional novel smORFs.	19	**22**/140 smORFs with evidence of trancription, or translation, or conservation produce growth deffects in different conditions.	15
Drosophila	Conservation, purifying selection, synteny and transcription	401	Putative functional novel smORFs.	17	***sarcolamban:*** regulates SERCA mediated-calcium uptake and heart rythmicity in *Drosophila.*	48
Arabidopsis	Hexamere sequence-composition frequencies, conservation, transcription and purifying selection	3,241	Putative functional novel smORFs.	16	**49**/473 conserved and transcribed smORFs produce over-expression morphological phenotypes.	47
Mouse	Hexamere sequence composition (CRITICA), and purifying selection	1,240	Putatively functional smORFs, of which 495 lack similatity to known proteins.	18	Epitope-tag of **14**/25 show translation and specific subcellular localisation in HeLa cells.	18
Drosophila S2 cells	RBF density and coverage	228	Out of 274 annotated smORFs.	32	Epitope-tag of **7 LncRNA** smORFs and **3 uORFS** show translation and specific subcellular localisation in S2 cells.	32
		313	Out of 918 possible smORFs in 125 lncRNAs transcribed in S2 cells.			
		2,708	Out of 9,069 possible uORFs within S2 cell transcripts.			
	PhastCons	212	Pass a threshold that distinguishes intergenic from coding regions at 10% FDR (out of 228 annotated, translated smORFs).		***hemotin:*** regulates endocytic maturation and phagocytosis in *Drosophila* hemocytes.	54
	Custom database	99	Out of 228 annotated smORFs, 60 found in this study, plus 39 from PeptideAtlas.			
Zebrafish embryos	**Enhanced Translated ORF Classifier (TOC).**	399	Novel coding genes (using chew et al ribo-seq data).	52	***Toddler:*** promotes cell migration during gastrulation by activating Apelin receptor signalling.	52
		89	smORFs (12 secreted and 15 TMHMM).			
Human cells	Protease inhibition, and ELRIC fractionation Custom database	86	Novel peptides,49 mapping to previously un-annotated transcripts. 8 mapping to LncRNAs,and 37 to 3'UTRs, 5'UTRs, and overlaping annotated CDSs.	44	***MRI-2:*** stimulates double-strand break repair through a direct interaction with the DNA end binding protein Ku.	56

Computational	Ribo-seq	Mass-spectrometry

BOX-1. Bioinformatic assessment of ORF-coding potential and translation.

sORF finder: bioinformatic package to identify smORFs with high confident coding potential based on their similarity in nucleotide composition to bona fide coding genes by hidden Markov model. Potential coding sORFs are further tested for functionality by searching homologues and evolutionary constrains [16].

Coding Region Identification Tool Invoking Comparative Analysis (CRITICA): gene prediction algorithm, which intergrates a purifying selection analysis of pair-wise aligned homologous regions into a hexamere sequence composition-analysis [18].

PhastCons: program that predicts conserved elements in multiple alignment sequences. It is based on a statistical hidden Markov phylogenetic model (phylo-HMM) that takes into account the probability of nucleotide substitutions at each site in a genome and how this probability changes from one site to the next [20].

PhyloCSF: comparative sequence method that analyses multiple alignments of nucleotide sequence using statistical comparison of phylogenetic codon models to ascertain the likelihood to be a conserved protein coding sequence [21].

Micropeptide detection pipeline (micPDP): method that evaluates the existence of purifying selection on aa sequence from codon nucleotide changes. This pipeline filters candidate alignments according to coverage and reading frame conservation and then PhyloCSF method is applied to assess their coding potential from codon substitutions in genome-wide multi-alignments [31].

ORFscore: translation-dependent metric that exploits the 3-nt step movement of translating ribosomes across the transcript. Therefore, the Ribo-seq reads in coding ORFs tend to show a tri-nucleotide codon periodicity on the frame of translation (phasing)[31]. This method requires a strict cut-off for the size of analysed RBFs (only precise size reads, usually 28-29nt, are used), which could lead to a significant loss in read density.

Ribosome Release Score (RRS): metric based on the releasing ability of ribosomes from the translating RNA after they encounter a stop codon. RRS is defined as the ratio between the total number of Ribo-seq reads in the ORF and the total number Ribo-seq reads in the subsequent 3'UTR, normalized respectively to the total length of their regions divided by the normalized number of RNA-seq reads in each region computed in the same fashion [78].

Fragment length organisation similarity score (FLOSS): this method relies on the difference of the fragment size distribution of the ribo-seq footprints in coding genes and non-coding RNAs. This metric scores the coding potential of ORFs according to their similarity of the length of protected footprints of known coding genes [58].

BOX-2. Evaluation of coding potential and translation of smORFs by computer learning classifiers.

Coding Potential Calculator (CPC): bioinformatics tool that scores six sequences features to distinguish coding vs non-coding ORFs, three relate to the quality of the longest ORF (ORF size, Coverage, integrity) whereas the other three are based on sequence conservation using BLASTX (number of hits, quality of the hits, frame distribution of hits) that are incorporated in a Support Vector learning machine classifier. [79-80].

Translated ORF Classifier (TOC): a Ribo-seq classifier based on a random Forest model that assess the coding potential of each ORF within a transcript based on 4 metrics: Translation Efficiency (ratio of the Ribo-seq reads/RNA-seq read within the ORF: Level of translation), Inside vs Outside (coverage inside ORF/coverage outside ORF; coverage is number of nucleotides having Ribo-seq reads/total number of nucleotides), Fraction Length (fraction of the transcript covered by ORF) and Disengagement score (DS) assess the release efficiency of the ribosome after a stop codon which is a characteristic of ribosome translating coding ORFs by measuring the Ribo-seq reads in the ORF/Ribo-seq reads downstream. [29]. Pauli *et al*.[52] improved TOC classifier by adding a "cover" metric (number of nucleotides of the ORF covered by Ribo-seq reads).

ORF Regression Algorithm for Translational Evaluation of RPFs (ORF-RATER): this metric is able to identify and quantify translation in ORFs from Ribo-seq data by comparing the patterns of ribosome occupancy (initiation and termination peaks and elongation phase) to that of coding ORFs. ORF-RATER uses a linear regression model that allows the integration of multiple lines of evidence and evaluates each ORF according to the nearby context [59].

RibORF Classifier: a Ribo-seq Support Vector Machine classifier that defines active translation of ORFs according to the evaluation of phasing by using 5' footprint off-set distances to the ribosome A-site, from canonical proteins, to identify 3nt periodicity, and uniformity of footprint distribution across codons by calculating the percentage of maximum entropy values [60].

Supplementary Table 1: Number smORFs identified using computational, Ribo-seq and proteomics approaches in different organisms.

Ref.	Organism:	Method/Metrics:	Indentified smORFs:	Comments:
Kessler *et al.* (2003)	**Yeast**	Conservation, transcription, optimal codon adaptation index	**588**	Putative functional novel smORFs.
Ladoukakis*et al.* (2011)	**Drosophila**	Conservation, purifying selection, synteny and transcription	**401**	Putative functional novel smORFs.
Hanada *et al.* (2007)	**Arabidopsis**	Hexamere sequence-composition frequencies, conservation, transcription and purifying selection	**3,241**	Putative functional novel smORFs.
Frith *et al.* (2006)	**Mouse**	Hexamere sequence composition (CRITICA), and purifying selection	**1,240**	Putatively functional smORFs, of which 495 lack similatity to known proteins.
Aspden *et al.* (2014)	**Drosophila S2 cells**	RBF density and coverage	**228**	Out of 274 annotated smORFs.
			313	Out of 918 possible smORFs in 125 lncRNAs transcribed in S2 cells.
			2,708	Out of 9,069 possible uORFs within S2 cell transcripts.
		PhastCons	**212**	Pass a threshold that distinguishes intergenic from coding regions at 10% FDR (out of 228 annotated, translated smORFs).
		Custom database	**99**	Out of 228 annotated smORFs, 60 found in this study, plus 39 from PeptideAtlas.
Bazzini *et al.* (2014)	**Zebrafish embryos**	RBF coverage and phasing (ORFscore)	**302**	Out of 581 annotated smORFs.
			190	Out of 2540 possible smORFs in un-annotated transcripts and LncRNAs.
			311	uORFs
		Custom database	**98**	Previously annotated smORFs.
			6	smORFs in un-annotated transcripts and LncRNAs.
			17	uORFs
		micPDP	**63**	Out of 15,674 ORFs in transcripts without annotated CDS (23/63 found by ribo-seq).
Pauli *et al.* (2014)	**Zebrafish embryos**	**Enhanced Translated ORF Classifier (TOC).**	**399**	Novel coding genes (using chew et al ribo-seq data).
			89	smORFs (12 secreted and 15 TMHMM).
Crappe *et al.* (2013)	**Mouse E14 mESCs cells**	sORFfinder +SVM	**1,100**	smORFs in ncRNAs
			27, 371	Intergenic smORFs
			23,127	Other smORFs (intronic, overlaping coding exons).
		RBF Coverage, and start codon pile-up with Harringtonine treated samples	**528**	smORFs in ncRNAs (401 pass sORFfinder +SVM filter).
			226	intergenic smORFs (89 pass sORFfinder +SVM filter)
Lee *et al.* (2012)	**Human HEK93 cells**	Global initiation site sequencing (GTI-seq) with LTM treated samples	**288**	smORFs (median size 18 aa) out of 5763 ncRNA transcripts (Refseq).
			6,729	uORFs in 3352 genes (mostly smORFs).
Ma *et al.* (2013)	**Human (cells / tissues)**	Protease inhibition, and ELRIC fractionation Custom database	**237**	Novel peptides, 80% mapping to previously un-annotated transcripts and 20% to 3'UTRs, 5'UTRs, and overlaping annotated CDSs.
Vanderperre *et al.* (2013)	**Human (cells / tissues)**	16 HPLC-MS datasets from peptide Atlas matched to custom database	**1,256**	Novel peptides within annotated transcripts, 3'UTRs, 5'UTRs, and overlaping annotated CDSs.
Slavoff *et al.* (2013)	**Human cells**	Protease inhibition, and ELRIC fractionation Custom database	**86**	Novel peptides,49 mapping to previously un-annotated transcripts. 8 mapping to LncRNAs,and 37 to 3'UTRs, 5'UTRs, and overlaping annotated CDSs.

Computational	Ribo-seq	Mass-spectrometry